Venues and estate sale prices of Savannah, Georgia

12-15-2019

1. Introduction

Savannah is the oldest city in the state of Georgia. It is also an industrial center and an important Atlantic seaport. Savannah attract many residents and tourist due to its notable historic buildings, rich and ground art culture, warm weather, and access to the Atlantic Ocean. Savannah has long and tropical summer while the winter is rather short and mild. In term of architecture, Savannah is known as the nation's largest historically restored urban area.

The goal of this report is to provide some information about neighborhoods of Savannah, Georgia. These information, while not trivial to find, are particularly useful for current residents or prospective residents who are thinking of residing in this city. For instance, city residents can use the information about venues and estates sale prices in the city to choose a living area that fit their budgets and their social living style. In addition, city leaders, who wants to have an updates about the city estates and venues to help better managing the city. Investors also can find the information provided here useful in choosing the location for the new business.

2. Data

In this report, we will focus only on two main type of information which are:

- current estates sale prices
- venues in the city

Additionally, choropleth map will also be used in this report. For this, the neighborhood boundaries geojson data was obtained from Savannah Area GIS Open Data at link: https://opendata.arcgis.com/datasets/f6124ade9eba4ecfa61a16c69431c8f8_10.geojson

	Price	Estate Type	Neighborhood	Beds	Bath	Sqft	Year Built	Heating	Cooling	Parking	Lot Size	Schools	School Grades	School Ratings
0	199000	Single Family	Laroche Park	3	2	1562	1956	Forced air	Central	7.0	8712	Low Elementary School	PK-11	2
1	154900	Single Family	Victory Heights	3	2	1370	1957	Heat pump	No Data	2.0	7840	Shuman Elementary School, Hubert Middle School...	PK-5, 6-8, 9-12	2, 2, 4
2	135000	Townhouse	Godley Station	2	2	1247	2006	Forced air	Central	0.0	1742	Godley Station School, Groves High School	PK-8, 9-12	5, 2
3	262078	Single Family	Godley Station	5	3	2888	2019	Heat pump	Central	0.0	8145	Godley Station School, Groves High School	PK-8, 9-12	5, 2
4	254900	Single Family	Godley Station	4	3	2680	2011	Other	Central	0.0	9583	Godley Station School, Groves High School	PK-8, 9-12	5, 2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
704	279900	Condo	The Village/Rio/Armstrong	3	3	1735	1982	Forced air	Central	2.0	0	Windsor Forest Elementary School, Southwest Mi...	PK-5, 6-8, 9-12	3, 4, 3
705	235000	Multi Family	None	8	4	3190	1978	Forced air	Central	0.0	0	Thunderbolt Elementary School, Myers Middle Sc...	PK-5, 6-8, 9-12	3, 3, 4
706	269000	Single Family	Godley Station	4	3	2218	2019	No Data	No Data	2.0	9191	Godley Station School, Groves High School	PK-8, 9-12	5, 2
707	29900	NaN	Brookview/Skidway Terrace/Parkview	0	0	9147	0	NaN	NaN	NaN	0	Isle Of Hope School, Johnson High School	PK-8, 9-12	4, 4
708	239900	Single Family	Highland Park	3	3	2040	2018	Forced air, Heat pump	Central	0.0	8624	Godley Station School, Groves High School	PK-8, 9-12	5, 2

*Table 1: Head and tail view of the dataframe of current **estates sale prices** in Savannah, Georgia.*

The current estates sale prices are obtained from the real-estate website Zillow.com. A procedure was written to scrape the website for all the current estates in Savannah being listed for sale. The desired information extracted are sale price, estate type, neighborhood, number of bedrooms, number of bathrooms, square feet, year built, heating, cooling, number of parking, lot size, nearby schools, school grades, and school ratings. Estates those are for rent will be ignored. Those estates we choose to extract are for sale, new construction, auction and foreclosed. Sample of the dataframe is shown in Table 1.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Georgetown	31.966999	-81.218094	Savannah Concrete Contractors	31.970030	-81.221072	Home Service
1	Wilshire Estates/Savannah Mall/Tranquilla Woods	31.987075	-81.154211	Redbox	31.982732	-81.152879	Video Store
2	Wilshire Estates/Savannah Mall/Tranquilla Woods	31.987075	-81.154211	Sbarro	31.983906	-81.152970	Pizza Place
3	Wilshire Estates/Savannah Mall/Tranquilla Woods	31.987075	-81.154211	Southside YMCA	31.986233	-81.152697	Gym / Fitness Center
4	Wilshire Estates/Savannah Mall/Tranquilla Woods	31.987075	-81.154211	Cloud9 Medical Spa	31.987287	-81.151505	Spa
...	...	...	...	...	...	...	...
787	Avalon/Oglethorpe Mall	32.003617	-81.118268	motherhood maternity	32.002861	-81.121669	Women's Store
788	Avalon/Oglethorpe Mall	32.003617	-81.118268	Havertys	32.003976	-81.122428	Furniture / Home Store
789	Avalon/Oglethorpe Mall	32.003617	-81.118268	Starship	32.000441	-81.121059	Adult Boutique
790	Avalon/Oglethorpe Mall	32.003617	-81.118268	David's Bridal	32.004173	-81.123009	Bridal Shop
791	Avalon/Oglethorpe Mall	32.003617	-81.118268	Riitz Salon	32.002399	-81.123176	Salon / Barbershop

Table 2: Head and tail view of the dataframe of venues in Savannah, Georgia.

Data about the city venues was obtained using Foursquare API. The search radius used was 500 meters from neighborhood center. A sample of the venues dataframe is shown in Table 2.

For current or prospective residents, estates sale prices information is useful for choosing the houses with adequate living space while still stay within their price range. Then information about venues give the residents an idea about social places density of different neighborhood in the city. This also aid in helping choosing their residing location. On the other hand, an investor might want to use the city venues information to choose the location of their new business which has less competition.

3. Methodology

After obtaining the necessary data, we now proceed to exploratory data analysis. First, we will map out the physical location of each neighborhood in Savannah, Georgia to have an idea of their locations. Second, the main data analysis we will be using is k-mean clustering. K-mean is a popular unsupervised machine learning algorithm. In this report, we will use k-mean to group similar data together and examine the hidden patterns. One important step of k-mean analysis is choosing the proper k value. To achieve that, we will use the elbow method. In particular, we will be looking at the distortion vs. k-value plot and inertia vs. k-value plot to determine the value of k. Distortion is calculated as the average of the squared distances from the cluster centers of the respective clusters. Typically, the Euclidean distance metric is used. Inertia is the sum of squared distances of samples to their closest cluster center.

For the city venues data, we will analyze the popularity of different type of venues in the city by looking at the venues density in each neighborhood. We will group the neighborhood using k-mean clustering. Choropleth map will be used to visualize the amount of venues in each neighborhood.

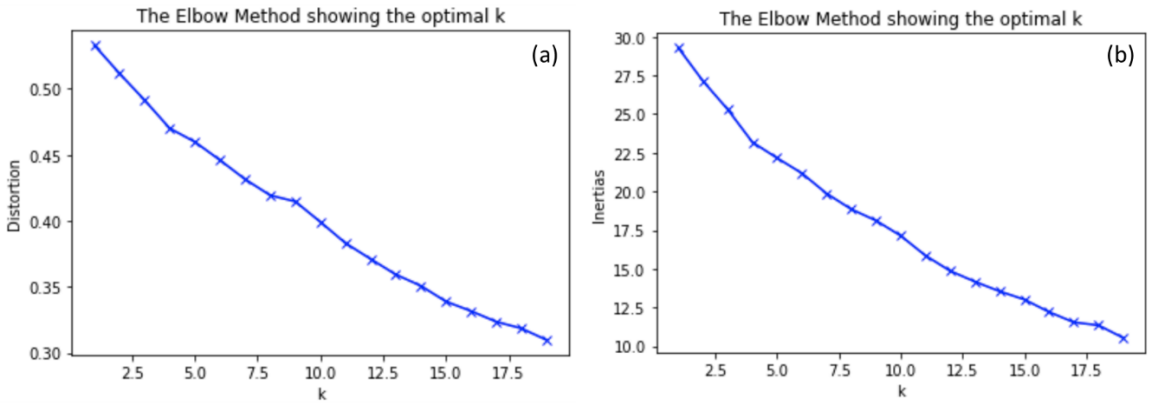

Figure 1. Determination of k value using elbow method on venue types density data. (a) Distortion plot. (b) Inertias plot.

From the above plots (Figure 1), it seems that there is no clear clustering in the city venues popularity data. Although the elbow is not obvious, we can see that there is evidence of an elbow in the inertias plot at k = 4. Therefore, I will choose k = 4 for further analysis.

We can examine the most common venue type in each neighborhood clusters to determine their characteristics

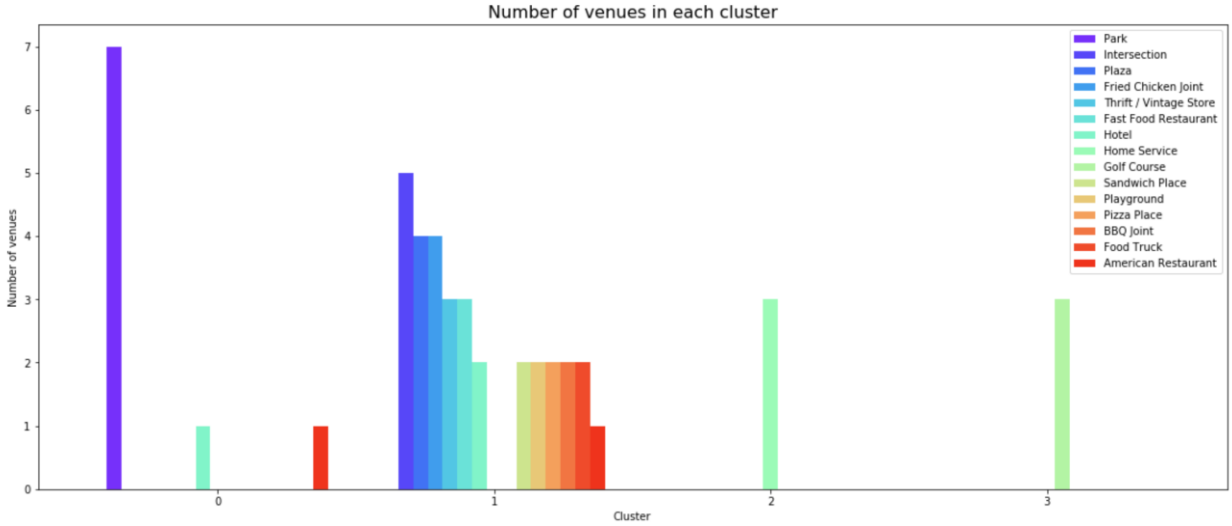

Figure 2: Distribution of the top 15 most common venue types in each neighborhood cluster.

From the above plot, each cluster of neighborhoods can be labeled as:

Clustered 0: parks.
Clustered 1: food and social venues.
Clustered 2: home service.
Clustered 3: Golf course

Similarly, we will analyze the estates sale prices in different neighborhoods in the city. An additional data filtering step is carried out. Estate listing which does not contain the require data (e.g. price,

number of beds, etc…) will be filtered out. We will first look at the average estates sale prices in each neighborhoods using bar chart, box plot, and heat map. We will attempt to cluster the neighborhood based on estate information (price, number of beds, bathrooms, square feet, year built).

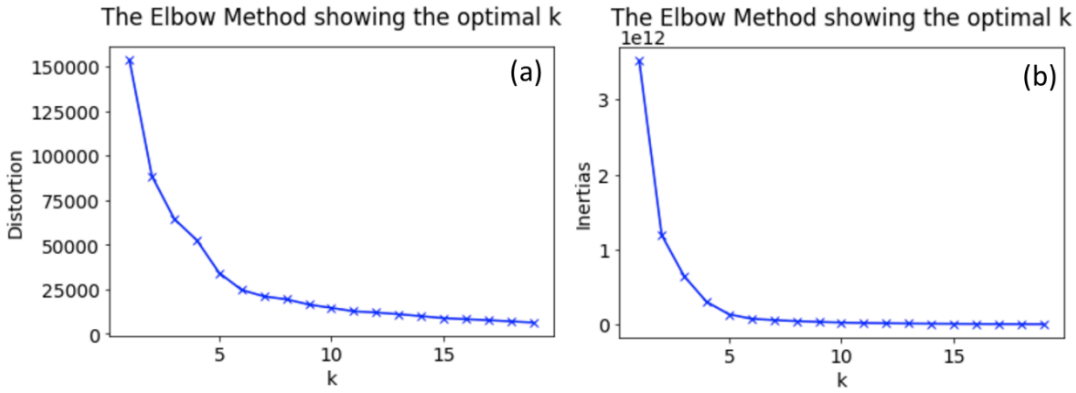

*Figure 3: Determination of k value using elbow method on **estates sale prices** data. (a) Distortion plot. (b) Inertias plot.*

From the Inertias plot (Figure 3), the optimal value of k is 5, which was then used for k-mean clustering.

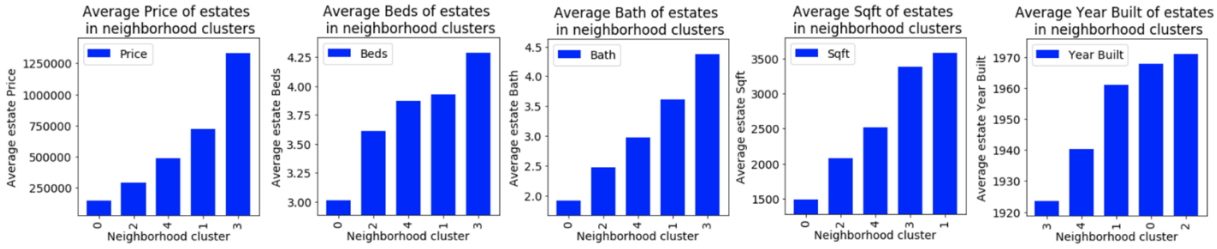

Figure 4: Selected estates' characteristic for different neighborhood clusters.

After the neighborhoods have been clustered, we need to find a description for each cluster (Figure 4). It appears that clustering falls into group of price, beds, bath and sqft, with cluster #3 being the most expensive, most number of beds, bath and largest sqft. These correlations don't seem to follow the year built. Year built follows a different clustering trend.

We can classify the neighborhood clusters as follow:
Cluster 0 = small newer estates with about 3 beds and sale price about 148K USD.
Cluster 2 = small-medium newer estates with about 3.6 beds and sale price about 291K USD.
Cluster 4 = medium older estates with about 3.9 beds and sale price about 485K USD.
Cluster 1 = medium-large estates with about 4 beds and sale price about 723K USD.
Cluster 3 = large older estates with about 4.3 beds and sale price about 1.3 million USD.

One important factor when choosing a residing location for a family with children is the quality of nearby schools. Therefore, we will examine to see if the ratings of nearby schools correlate with estates sale prices.

4. Results and discussion

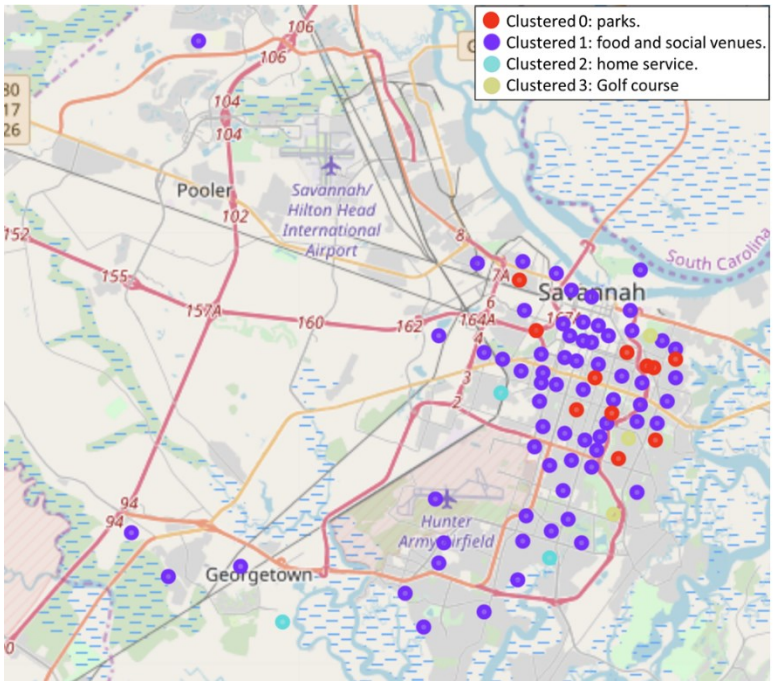

Figure 5: Map of neighborhoods in different groups based on venues popularity. The value of k=4 was used.

Regarding the city venues, our k-mean clustering analysis shows that there is only a weak evidence of clustering in the data set based on the distortion and inertias plots. That is the different type of venues seems to distribute rather evenly within the city instead of segregating in different areas. Most of the venues in the city are social and food venues which are distributed evenly. Nevertheless, there seems a be a weak "elbow" at k = 4, which I used for the clustering procedure. The clustered neighborhood based on venues are shown in Figure 5. While food and social venues distribute evenly within the city, those who prefer park and recreations might like to reside toward the north and north-east side of the city. Golf enthusiast can find a few golf courts on the east side of the city. The lack of venues clustering is understandable since Savannah, GA is not a very big city. It has a total population of about 146,000, which is magnitude less than the population of big city such as Toronto (2.8 million) and Chicago (2.7 million). Therefore, the amount of venues in Savannah is also a lot less, which leads to a lesser degree of business competition and segregation.

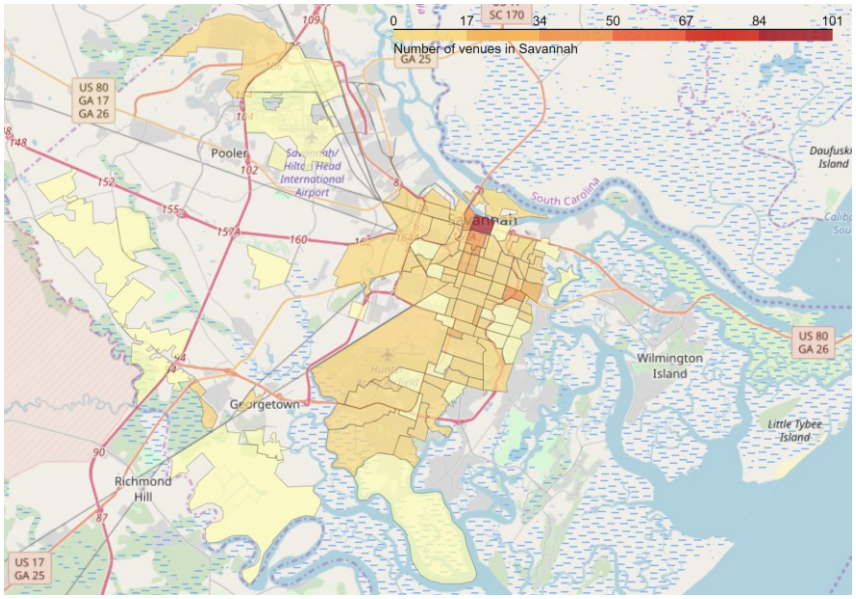

Figure 6: Choropleth map of total number of venues in each neighborhood.

Next, we looked at the total number of venues in each neighborhood. As expected, the number of venues is most densely populated in downtown Savannah, that is the South Historic District and North Historic District neighborhoods (Figure 6).

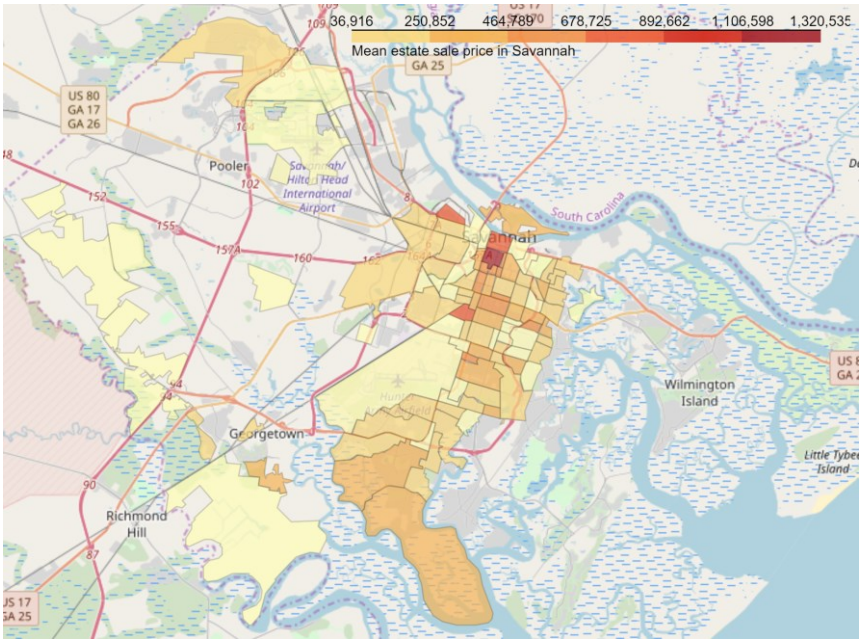

Figure 7: Choropleth map of the average current estates sale prices in each neighborhood.

For current estates sale prices, we first look examine the average prices in each neighborhood using choropleth map (Figure 7). The most expensive estates seem to locate in the downtown area (South and North Historic District), with an average sale price of 1.3 million USD.

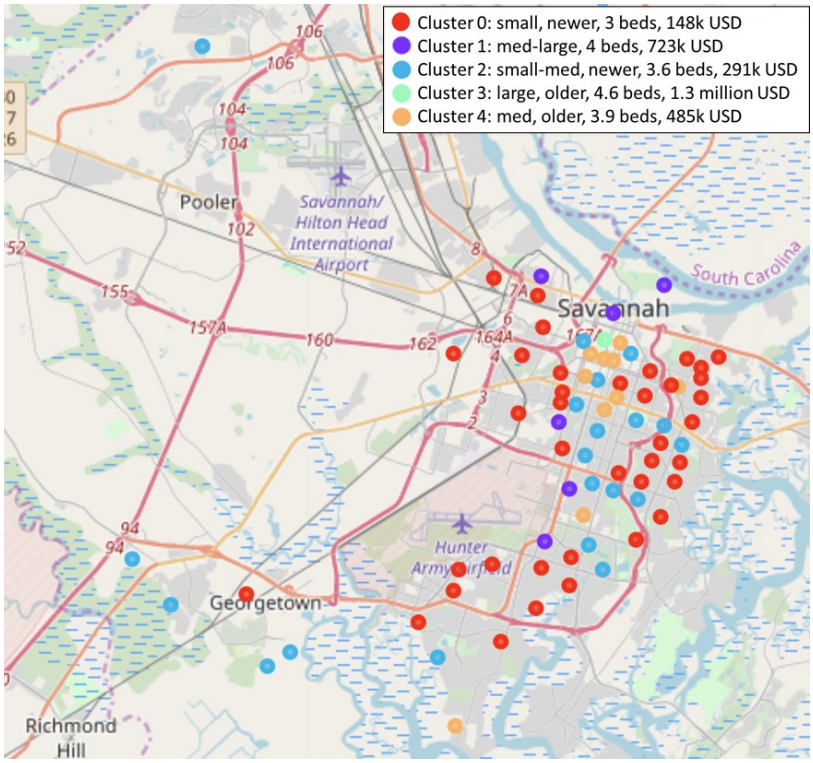

Figure 8: Map of neighborhoods in different groups based on selected estates' characteristics. The value of k=5 was used.

Interestingly, estates sale prices analysis using k-mean clustering showed a clear evidence of clustering with k = 5 (Figure 3). The neighborhoods segregate into 5 different groups with different range of estates sale prices and estate sizes. The most expensive, largest, and oldest estates locate in the downtown area in the South Historic District (Figure 8). On the other hands, newer estates were built away from the downtown area and are generally smaller and cheaper in price. As expected, estates sale prices appear to correlate linearly with the number of bedrooms, bathrooms, and square feet (Figure 4).

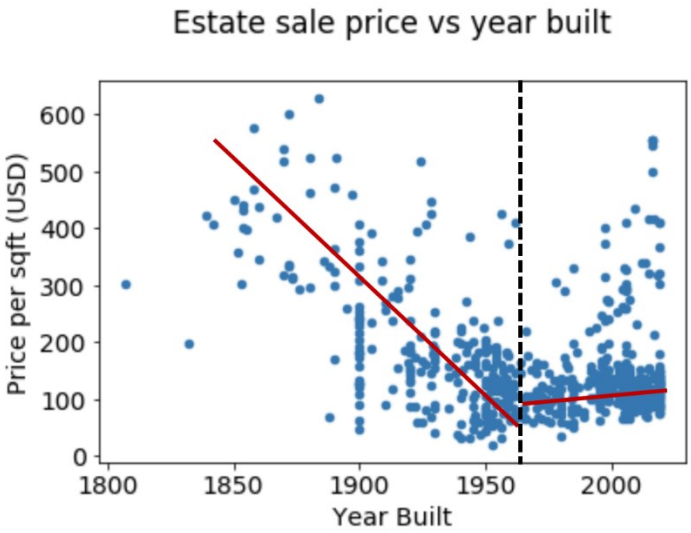

Figure 9: Sale prices per sqft for estates with different year built.

Interestingly, from the price_per_sqft vs year built plot (Figure 9) we see that there are two different trends present. The first trend is that estates built before 1960 appears to have decreasing sale prices for newer estates (up to about 1960). The second trend is for estates built from 1960 to present. In this case, the estates sale prices per sqft remains rather flat.

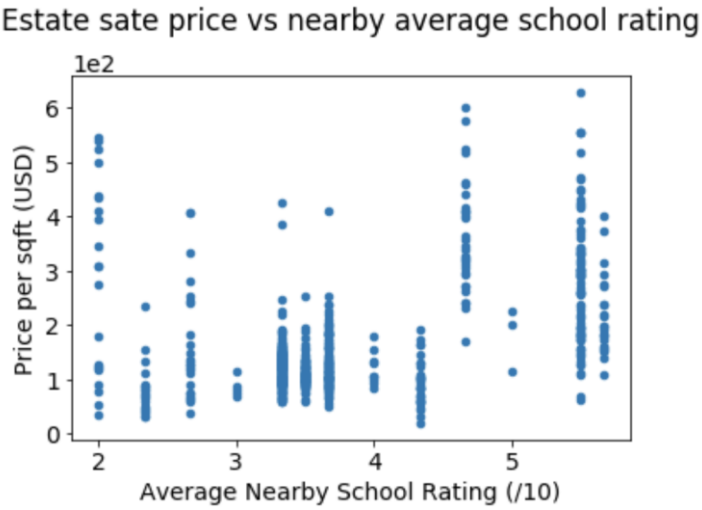

Figure 10: Sale prices per sqft of estates as a function of their nearby school ratings.

On the other hand, there seems to be no correlation between estates sale prices and the nearby school ratings (Figure 10). This means families can choose to settle in an area with lower house cost but will have good schools nearby for their children.

5. Conclusion

This report provides an analysis of venues and estates sale prices in the city of Savannah, Georgia. Current and future city residents can use this information helps aid them in making a good decision about where to live which suits their price range, live style. In addition, city leaders can use this information to aid in managing their city more effectively.